Cost-effective GPT inference accelerator using AiM (SK Hynix PIM)

Euicheol Lim

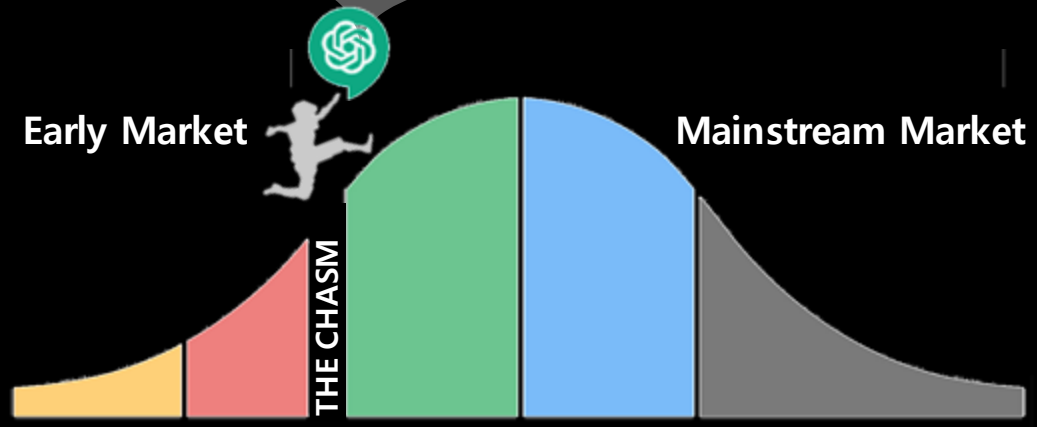
SK hynix

ChatGPT : Game changer of AI Market

- ChatGPT is opening a new mainstream market for AI Services but OpEx issues have to be solved

Innovative Technology | Investment | WOW Factor

ChatGPT Plus | GPT-4 | Microsoft 365 Copilot | ChatGPT Plugins



"ChatGPT is crossing the CHASM"

Operating Expenditure ✓

ChatGPT inference – Encoding & Decoding

- ChatGPT Inference consists of input processing (encoding) and answer generating (decoding)
- Especially, decoding is significantly memory intensive

Encoding : Comprehension

Input: 9 Words / Model Parameter Read: 1

h What are the issues of running ChatGPT with GPUs?

What	are	the
issues	of	running
ChatGPT	with	GPUs

Computing-Intensive

Decoding: Generating Answer

Output: 196 Words / Model Parameter Read : 261

Some potential issues in using GPT systems building ChatGPT systems are:

Cost : Large-scale models may require a significant number of GPUs, which can result in substantial costs. Especially, the latest high-performance GPUs can be very expensive ...

Power consumption : GPU consume ...

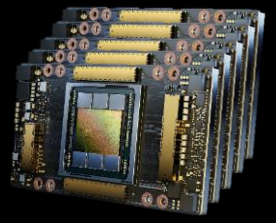
Memory-Intensive

1) For easy understanding, we use 'Words' as a unit of sequence length as opposed to 'Token'. Since 1 Token is the same as 0.75 Word on average, the number of 'Read' is 261.

Is GPU sufficient for ChatGPT inference decoding?

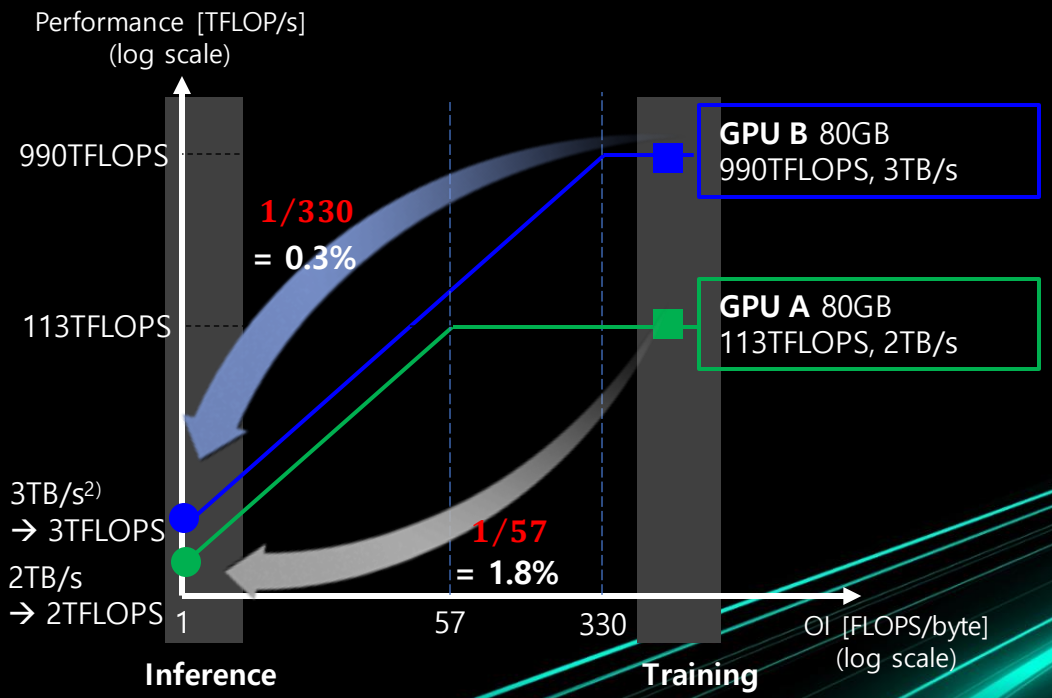
- GPU is not the cost-effective computing infra for chatGPT inference decoding

Inference Time with GPU System[1)]



	GPU A (80GB, 2TB/s) x 5	GPU B (80GB, 3TB/s) x 5
Model size	350GB	350GB
Bandwidth	10TB/s	15TB/s
Processing Time (1 token)	35 ms (350GB/10TB/s)	23 ms (350GB/15TB/s)
Processing Time (261 token)	9.1 sec	6.0 sec

Why so slow?

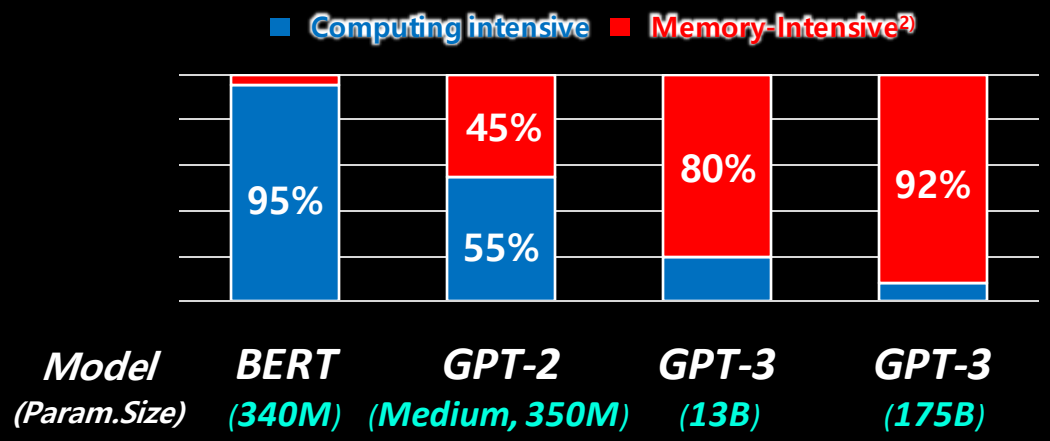


- **Extremely low performance per cost**
  - Using only under 2% of the GPU's performance, Wasting GPU power consumption

Why PIM as a GPT inference Accelerator??

- PIM is the best option for GPT inference decoding which is highly memory-Intensive.

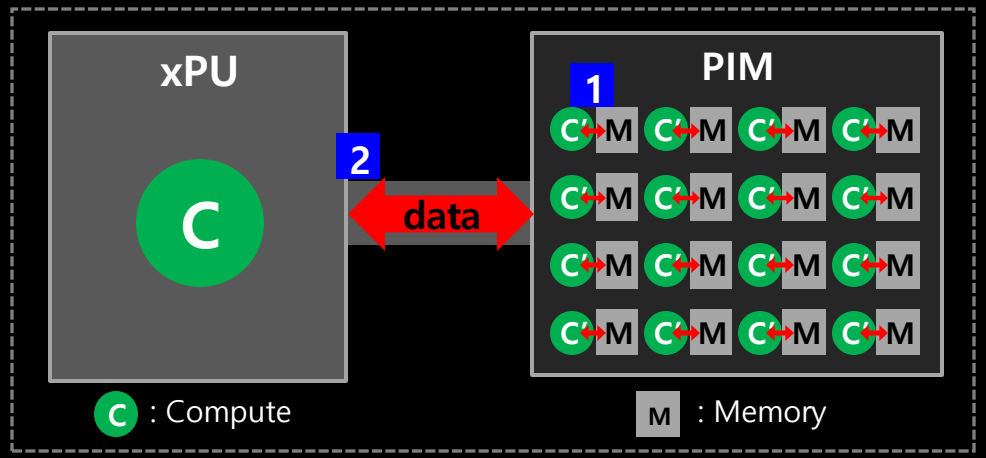
Feature of GPT Inference with Model & Size[1)]



- The larger the model, the **more memory intensive function (specifically, "GEMV"),** so **Memory Bandwidth for GEMV operation** has a greater impact on system performance than the processor

1) Measured data using 1x V100 GPU with PyTorch (v2.0)
2) Proportion of feed-forward network (FFN) which consists of Linear layer (GEMV, Matrix-Vector multiplication)

Feature of PIM



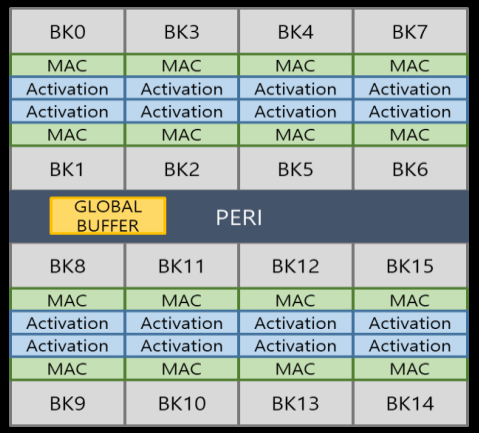
1 Performance Improvement
By utilizing the higher Bandwidth inside the memory

2 Energy Efficiency Improvement
By minimizing data movement between host and memory

PIM is suitable for accelerating Memory-Intensive Application like GPT inference

AiM introduction

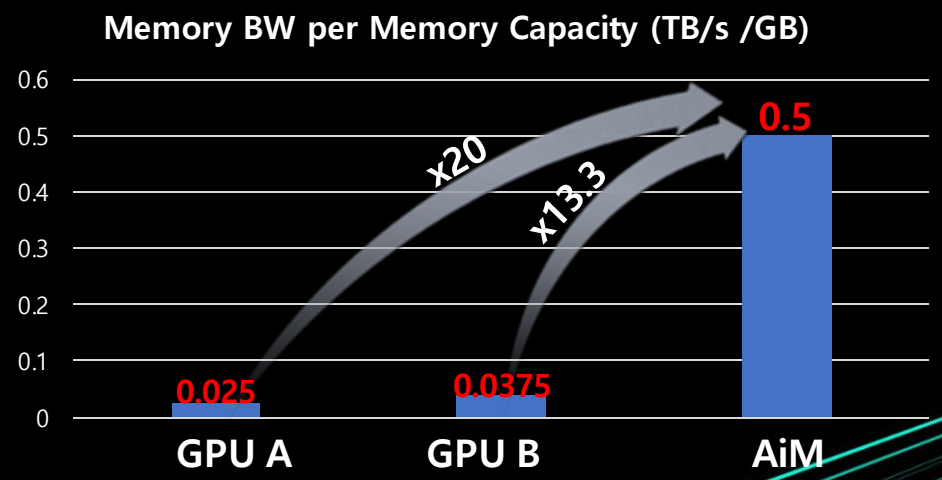
- SK hynix's very first GDDR6-based processing-in-memory (PIM) product called AiM(Accelerator-in-Memory) is ready and focused on GEMV operation.



AiM	
Memory Density	1GB
Bandwidth-external	64 GB/s
Function Support	GEMV, Activation
GEMV Bandwidth	0.5 TB/s (x8 of external BW)
GEMV Performance	0.5 TFLOPS
Numeric Precision	Brain Floating Point 16 (BF16)
Targets	Memory-intensive **GPT** applications

Critical Metric

Performance per Memory Capacity [TFLOPS/GB]
→ Memory BW per Memory Capacity [TB/s / GB] (if OI = 1)



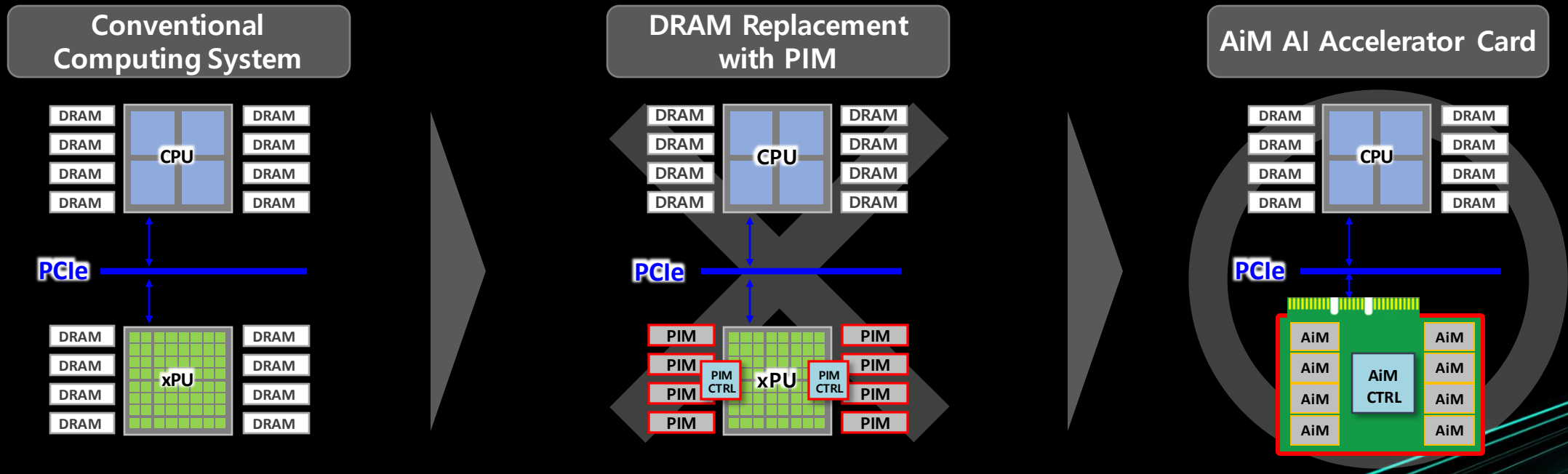
AiM outperforms GPU A, GPU B in terms of Performance per Memory Capacity [TB/s /GB = TFLOPS/GB]

$$\text{Token processing time} = \frac{1}{\text{Critical Metric}}$$

SK hynix

How to deploy AiM into existing system

- Can be easily deployed into the existing system by simply adding the AiM based AI Accelerator Card, rather than by replacing DRAM with PIM



- Memory bottleneck

- Host SoC (xPU) must be modified[1)]
- SW burden for memory mgmt.

- No need to modify any existing xPUs
- Need an AiM controller chip
- SW modification can be minimized

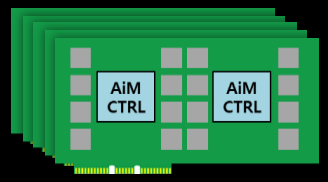
1) Conventional memory controller + additional command for PIM operation + in order scheduling

AiM based accelerator system for GPT service

- An AiM based GPT accelerator card system has 13 times performance than GPU B card system

Inference Time with AiM System[1)]

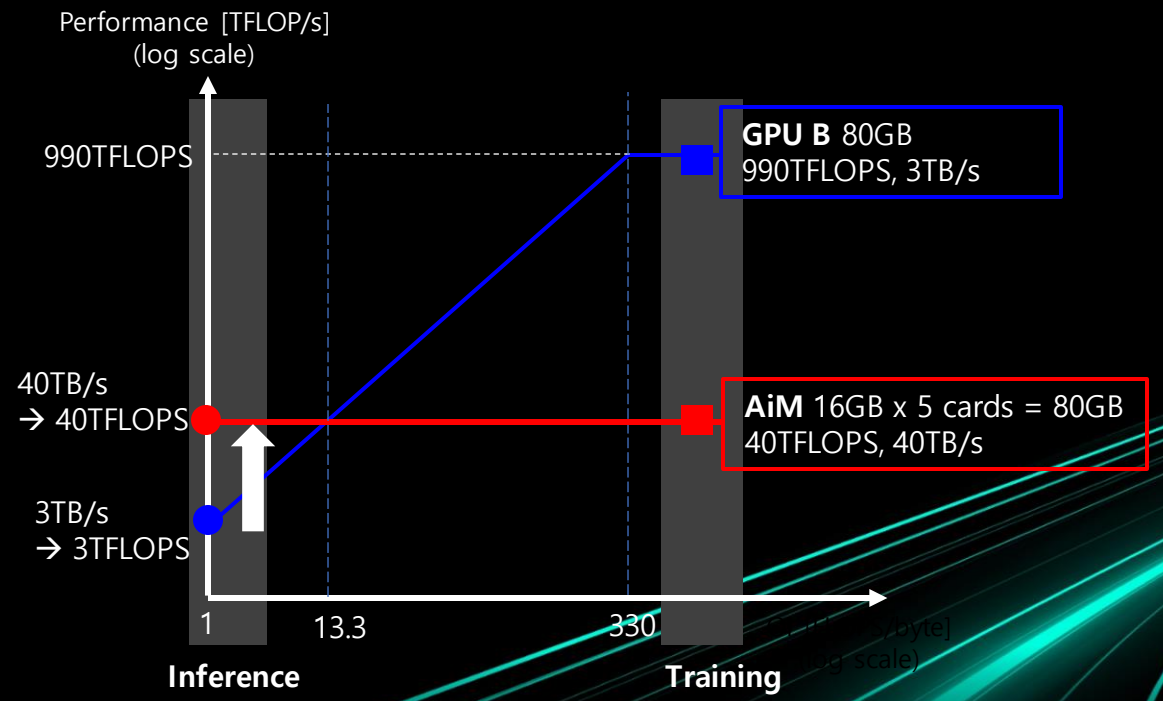
GPU B (80GB, 3TB/s) x 5



AiM (16GB, 8TB/s) x 25

Model size	350GB	350GB
Bandwidth	15TB/s	200TB/s
Processing Time (1 token)	23 msec (350GB/15TB/s)	1.8 msec (350GB/200TB/s)
Processing Time (261 token)	6.0 sec →	0.46 sec

Performance Comparison



13 times shorter latency means that 13 times as many services can be done per unit time

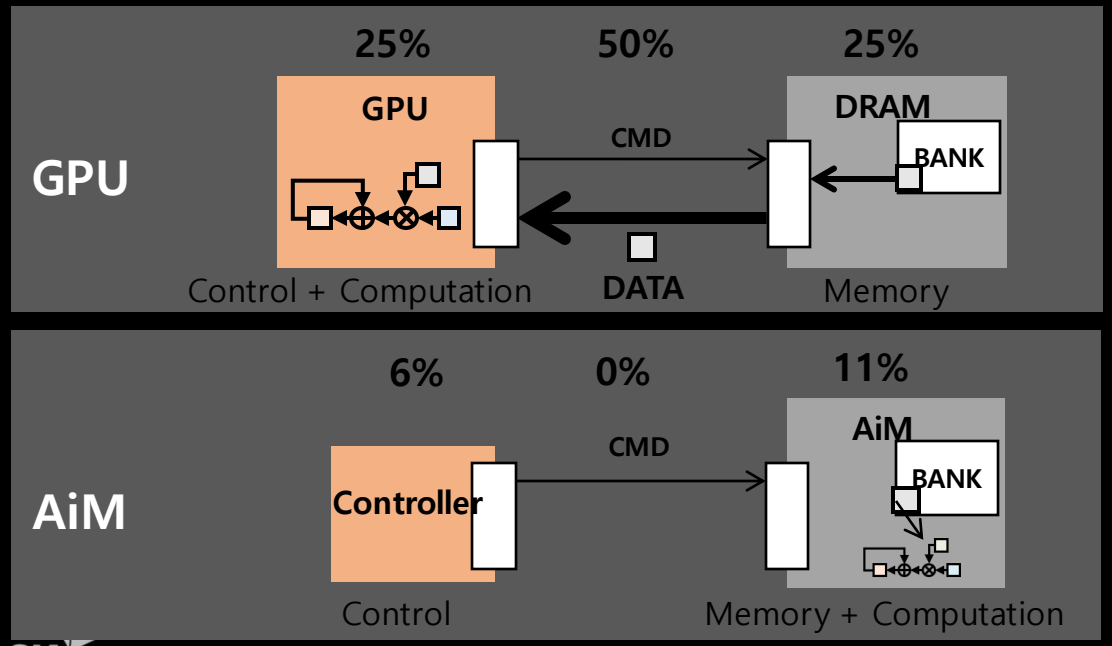
SK hynix

AiM based system – Energy consumption

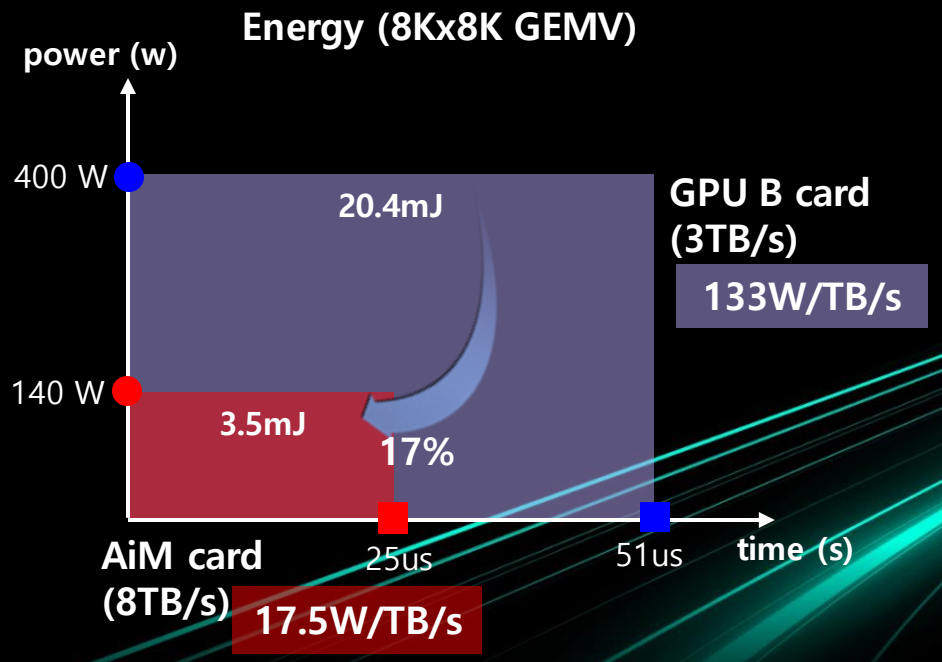
- AiM based system's energy consumption for GPT service is reduced to around 17% of GPU

GEMV energy consumption : 17% of GPU

1. Small controller and dedicated MAC unit in AiM
2. Remove off-chip data movement and reduce internal data movement
3. Reduce the static energy by short processing time



Critical Power Metric



AiM Benefit Summary

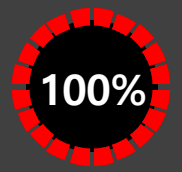
13x performance improvement is expected with 17% energy consumption and 50% system price compared to GPU B system for the GPT inference

GPU B system: $160,000

GPU B (80GB) x 5

Server x 1

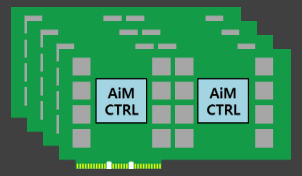
Performance



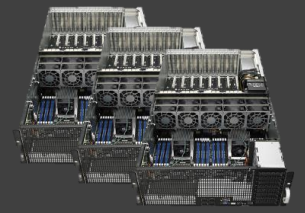
Energy Consumption

Price

AiM based system: $80,000

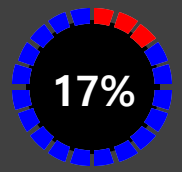


AiM Accelerator Card(16GB) x 25



Server x 3

Performance



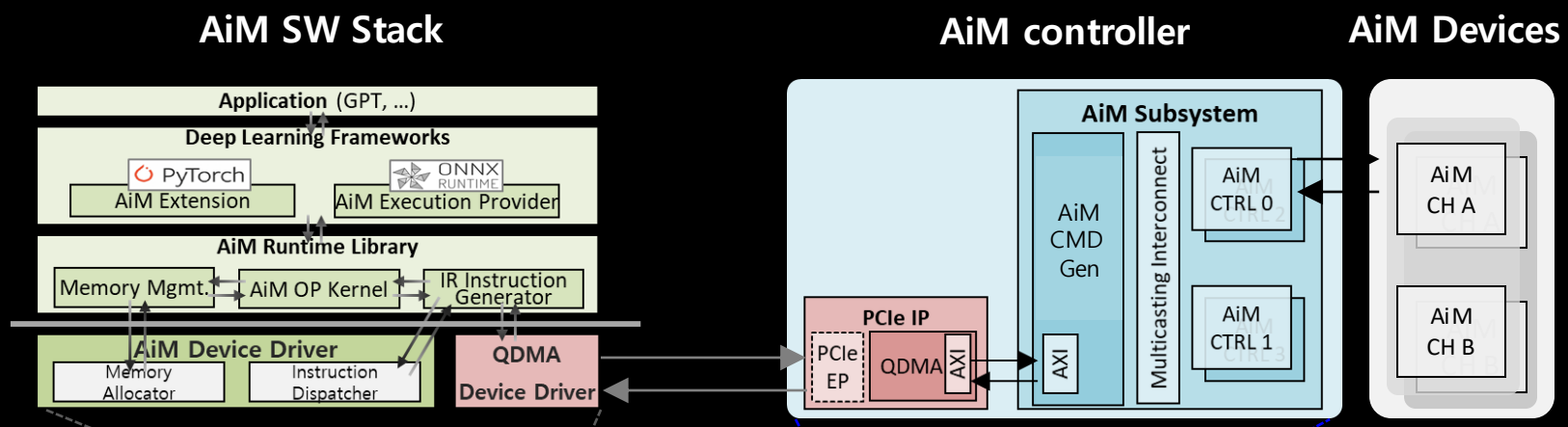
Energy Consumption

Price

- **Assuming serving GPT-3 (175B) model with single batch – Minimum 350GB required**

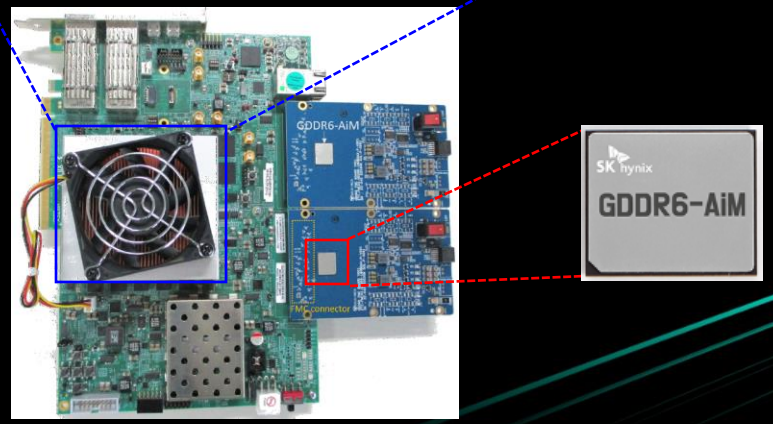
Readiness of AiM solution

- Completed the development of evaluation platform by connecting 2 AiM chips to commercial FPGA boards and finished the demo for sentence generation by porting of GPT2



Form Factor	Commodity FPGA Evaluation Board
Card Spec	1 CTRL x 2 chips @ 2Gbps
GEMV BW	0.125TB/s
Capacity	2GB

X86/ARM Server



Evaluation Board
(FPGA + daughter board with AiM chip)

[AiM Demonstration] GPT-2 Text Generation

SK hynix

Configuration

Input Prompt: Text Field

Generate Length: Text Field

Generate

Text Generation Result

Clear

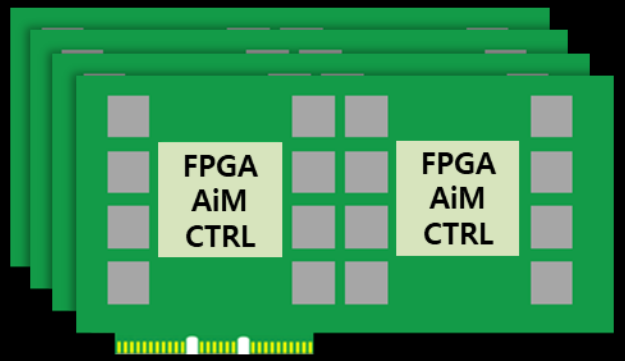
Performance Result

GPT-2 Medium Model

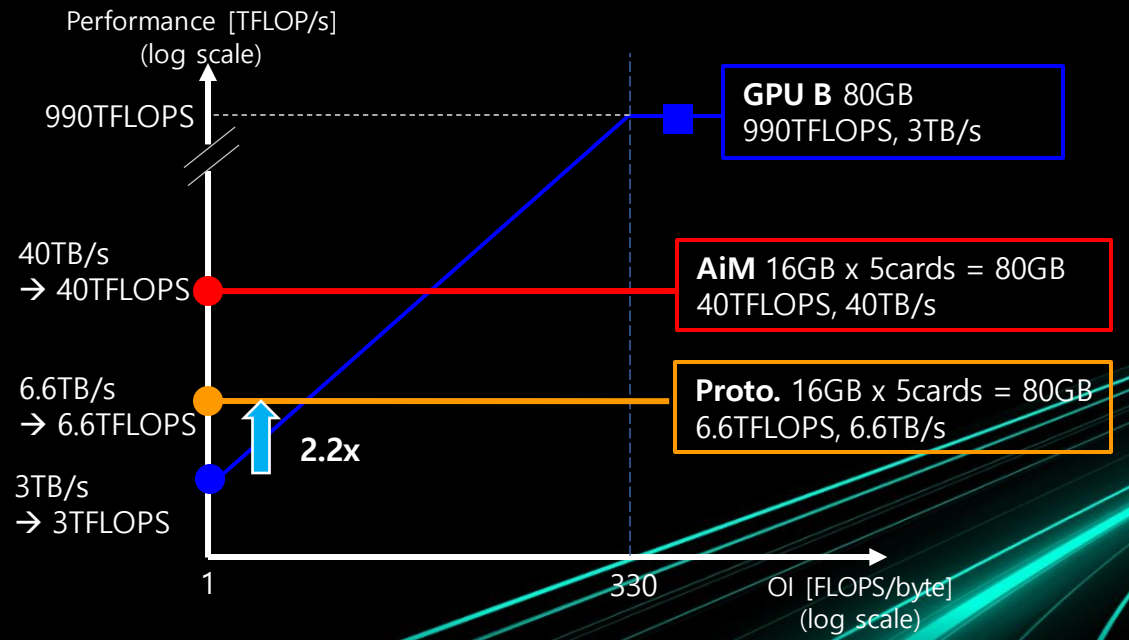
Reference Prototype under development

- We are developing a prototype card using FPGA chip as a reference design, which provides 1/6 of performance than ASIC AiM controller case, but still has better performance than GPU B
- Plan to have a showcase at AI hardware summit.

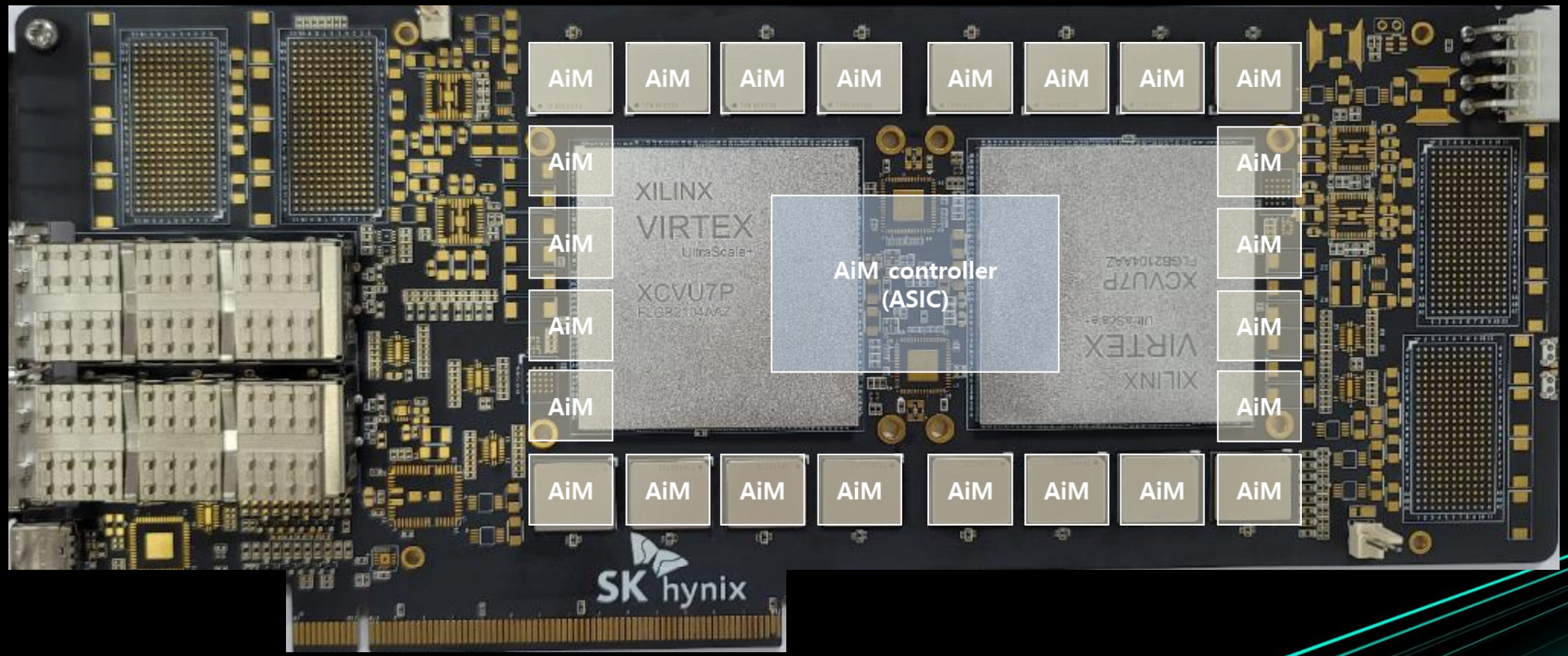
Form Factor	FHFL (A100/A30 compatible)
Card Spec	2 CTRL x 16 AiM chips @ 2.6Gbps
GEMV BW	1.33 TB/s
Capacity	16GB
Scale out	Chip2chip comm.



AiM AI Accelerator Card Prototype



Reference prototype card



SKhynix_PIM@skhynix.com

Summary

What are the issues of running ChatGPT with GPUs?

Some potential issues in using GPT systems building ChatGPT systems are:

Can be improved by AiM

Cost: Large-scale models may require a significant number of GPUs, which can result in substantial costs. Especially, the latest high-performance GPUs can be very expensive, making it costly to build a system using them.

Power consumption: GPUs consume a significant amount of power, requiring adequate power supply to operate them. Moreover, operating a large GPU cluster can consume a lot of power, increasing the cost of electricity.

Cooling: GPUs generate a lot of heat when processing large-scale models. If this heat is not removed, GPUs can overheat and be damaged. Therefore, an appropriate cooling system is necessary.

Bandwidth: Processing large-scale models requires the transfer of a significant amount of data. Thus, high bandwidth is necessary, requiring a network infrastructure and distributed file system that supports it.

Data management and preprocessing: Large-scale models require a lot of data to be trained and properly preprocessed. This requires a suitable data management and preprocessing system.

Scalability: Processing large-scale models requires a highly scalable system. However, GPU-based systems may face challenges in achieving optimal scalability due to the limited number of GPUs that can be integrated into a system.

SK hynix